International Journal of Basic and Applied Biology
p-ISSN: 2394-5820, e-ISSN: 2349-2539, Volume 4, Issue 1; January-March, 2017, pp. 22-26
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/Publication.html

Evidence and Positive Selection Test for Gene-loss Theory of Evolution

Prateek Sukumar[1], Neha Kumari[2] and Yasha Hasija[3]

[1,2,3]Department of Biotechnology, Delhi Technological University, Shahbad Daulatpur, Main Bawana Road, Delhi-110042, INDIA
E-mail: [1]prateek.dtu@hotmail.com, [2]sngh367@gmail.com, [3]yashahasija@gmail.com

Abstract—*The standard theory of evolution by Darwin attributes the development of all complexities manifested by different life forms to forces of natural selection and survival of the fittest. It also states that the need for adaptation to survive and produce off springs has caused the life to develop from low complexity single cell organism to multi cellular highly complex Human. In the present study, we argue that if in fact natural selection shapes different forms of life then life must have devolved from higher complex form to the lower complex form. The major premise being losing complexity improves your chances at survival. We try to prove it through documenting various kinds of evidences through examples. Moreover an experiment has been carried out in which we determine the positive selection between 420 pairs of homologous genes between various mammalian species. The genes were tested using Codon based z -test of positive selection by forming a pipeline through MEGA software. In 83% of total sequence pairs positive selection was not present. This suggests that our genes are not under positive selection as suggested by Darwin theory of evolution.*

1. INTRODUCTION

Origin of Species is one of the most difficult question faced by mankind. Darwin theory of evolution is most widely accepted theory in this regard. Darwin explained it using morphological features. But with further insights at level of nucleotides, several scientist have started to raise questions on plausibility of formation of species according to theory of evolution. Motoo Kimura came forward with neutral theory of evolution. And more recently "Gene loss" theory of evolution is becoming more popular.

Gene loss theory or "less is more" hypothesis is coined by Maynard V. Olson. Through this hypothesis he proposes a testable view that gene loss is a major motif of molecular evolution [1]. This is carried out by taking several examples from mice, yeast and human genomes that support adaptation and survival through gene loss.

This hypothesis is made concrete by observing the difference between the reproductive behavior of wild and laboratory mice. Wild strains of *Mus musculus*, the species from which the laboratory mice were derived, show a seasonal manner or pattern of reproduction. These wild strains show the similar diurnal cycles of melatonin synthesis that occur in the pineal gland as shown nearly by all mammals that have central role in the regulation of seasonal reproduction as described by Tamarkin [2]. This mechanism is evolutionary conserved, this mechanism monitors the changes in the length of daylight and adjust the reproductive behavior accordingly in that response. As this is evolutionary conserved mechanism it must be the ancestral state. But when we study the laboratory mice whose reproduction is uncoupled from seasonal change it now show know pineal-melatonin synthesis. This feature is because of the occurrence of recessive mutations in two genes, which code for the two enzyme required for the conversion of serotonin to melatonin [3,4]. A plausible hypothesis is that these mutations occur due to selection for trait of unregulated breeding, a highly desirable characteristic of domesticated mice.

In Humans, the instances of adaptive gene loss included Duffy-negative blood group and its relationship with resistance towards *Plasmodium vivax* [5]. This involves loss of chemokine receptor that are essential for entry of the pathogens into target cells. In this example, there is the occurrence of recessive mutation in a promoter element required for expressing the receptor DARC in erythroid lineages. In some regions of western Africa, there is 100% allele frequency of Duffy-negative mutation.

The strength of this hypothesis is that it can be readily tested apart from the fact its genetic plausibility. The testing of this loss-is-more hypothesis rests on the ease with which gene loss mutations could be easily recognized through sequence analysis. Many human genetic diseases, as phenylketonuria, cystic fibrosis, some types of breast cancer 1 that involve early onset of BRAC1 and BRAC2 genes, seem simply to require loss of the relevant gene function [6]. Existing data that proves less is more hypothesis are indeed good but are limited. The best described differences between Homo sapiens and Pan troglodytes adhere to this hypothesis. The example of it is in one major biochemical difference that humans cannot synthesize a form of the cell-surface SALIC ACID called N-glycolyl-neuramininc acid [7]. A study has been conducted that checks whether there is process of adaptive pseudogenization involved with human origin [8]. The adaptive pseudogenization would actually mean selection by loss of genes. A comparative genomic analysis was carried out to identify 80 non processed pseudogenes that got inactivated in Homo sapiens after its separation from chimpanzee lineage. The functions involving chemoreception and immune response were found over represented. However to study adaptive pseudogenization the focus was on CASPASE12 gene, a cysteinyl aspartate proteinase that participates in inflammatory and innate immune response to endotoxin. Through population genetic evidence it has been found that there is nearly complete fixation of a null allele at CASPASE12. And this process is being driven by positive selection as null allele would provide protection against sepsis. It was also estimated that pseudogenization of CASPASE12 started shortly after out-of-Africa migration of Humans. Interestingly, two more genes that were also associated with sepsis were also pseudogenized in humans. Thus the identification and analysis of human –specific pseudogenes open the door for understanding the roles of gene losses in human origins, and the finding that gene loss is in itself an adaption that supports “less-is-more” hypothesis.

The gene loss hypothesis is particularly more intriguing in human evolution, many gene losses have been proposed that provide adaptations and are responsible for specific human phenotypes. As, the pseudogenization of MYH16 in human is responsible for reduction in the size of hominin masticatory muscles that provided space for brain size expansion; it is fascinating to identify and analyze all of the human-specific gene losses i.e. the gene losses that occurred after the human-chimpanzee divergence event. In this study this would the gene loss may have occurred independently in other species also except chimpanzee. First step is the identification of human – specific gene losses by the comparison of human nonprocessed pseudogenes with the chimpanzee genome sequence. Such human specific pseudogenes were formed after the separation event of human – chimp in the last 6-7 million years [9]. The genome of human has abundance of pseudogenes [10] but most of them are found to be processed. Processed pseudogenes are DNA sequences that are reversed transcribed and then randomly inserted in the genome [11]. Such genes never had any function and hence eliminated from the study. In contrast, nonprocessed pseudogenes are those sequence that once had a function but now have their coding sequence interrupted. But many of these non processed pseudogenes are formed after event of gene duplication to avoid genetic redundancy [12].

Darwin theory of evolution seems quite plausible at morphological level. The gaining of new morphology or phenotype by addition of slight variations in phenotype of the species looks reasonable. But a further inquiry at genetic or protein level opens the huge complexity involved that would be required to gain a new characteristic. It has been shown that the vast majority of possible protein sequences formed for a given size shall be unstable and cannot be maintained inside the body of living organism. Let apart being functionally beneficial [13, 14, 15, 16]. Take a comparatively small protein sequence of 150 aa in length. The total number of sequences that can be formed by random permutations and combinations shall be 20^{150}. Out of these total possible structures how many protein sequence can possibly produce a stable protein, let apart a beneficial function. Experimental studies have shown that only 1 in 10^{74} sequences of 150aa in six are capable for the formation of a stable protein and this ratio decreases exponentially with increase in size of the protein.[17, 18]. To produce a functional protein the ration becomes 1 in 10^{77} protein sequences.

Darwin himself acknowledged that - "If it could be demonstrated that any complex organ existed which could not possibly have been formed by numerous, successive, slight modifications, my theory would absolutely break down." In this light comes the concept of irreducible complexity [19]. By irreducible complexity it is meant that there is a single system that is formed or composed of several interacting parts that together contribute to a single basic function and the removal of any one of the part shall cause the cease of function of the system.

There are as such many examples of biological complex system but most famous is the bacteria flagellar motility system. It consists of 50 genes that include genes for the sensory apparatus that turns flagellum clockwise or anti clockwise according to the environment and 40 other structural genes that builds the whole flagellum. The DNA required for building of flagellum is over 10000 codons [20]. And this requirement is irreducible in order to attain the function of flagellar motility. The argument is that how can nature bring all the parts together gradually when the system does not work until all the parts are in their unique place at the same time.

There are also human specific examples where loss of genes in humans help them to survive. Activation of EGFR related pathways induces genes implicated in tumor progression. While removal or blocking of these pathways would mean better survival as this would prevent cancer to occur. In fact we already have in place the EGFR targeted therapies(eg. Cetuximab for colon cancer) block the activation of this signaling pathway [21]. Clinicians have noticed that a small fraction of people engaged in high risk behavior did not develop AIIDS. The reason deduced is the functionality of co receptor of CD4 being lost. As HIV binds to CD4 and its co-receptor, the loss of co-receptor would mean that HIV won’t be able to bind and hence enter inside the cell to cause infection [22,23]. The insertion sequence IS6110 has been associated with new resistance emerging through the inactivation of critical genes. Transposition of the insertion sequence IS6110, was identified in the pncA gene from 19 pyrazinamide-resistant Mycobacterium tuberculosis strains. Alignment of the PncA protein from homologous proteins from different bacteria species revealed three highly conserved regions in PncA which play an important role in the processing of pyrazinamide [24]. It study found that there is 40% lower risk of coronary heart disease in the people who carried the mutations that crippled the activity of APOC3 gene. Hence, suggesting inhibition of APOC3 as a new potential strategy for therapeutic development [25, 26].

2. MATERIAL AND METHODS

2.1 MEGA:

Availability: http://www.megasoftware.net/

The molecular evolution genetics analysis (MEGA) tool is an integrated software or suite that performs statistical based test on comparative analysis of molecular sequence data and is based on the evolutionary concepts [27, 28]. For high throughput analysis, MEGA has re-engineered its source code i.e. computational core – which implements the algorithms for all analysis in MEGA, and now provides it as an stand alone program. And can be executed through command line by use of scripting language as perl.

MEGA-CC comes with MEGA-Proto. MEGA – Proto allows the generation of a configuration file. It ask the user to provide parameters for the analysis and it is the mirror of the GUI based MEGA application. As soon as the user is done with the selection of parameters and configuration, this selection is saved as the file. Once this is done we can now execute MEGA-CC to perform our required task.

Test Hypothesis

One way to test whether positive selection is operating on a gene is to compare the relative abundance of synonymous and nonsynonymous substitutions within the gene sequences. For a pair of sequences, this is done by first estimating the number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N), and their variances: Var(d_S) and Var(d_N), respectively. Through this information, we test the null hypothesis that H_0: $d_N = d_S$ using a Z-test:

$Z = (d_N - d_S) / \mathrm{SQRT}(\mathrm{Var}(d_S) + \mathrm{Var}(d_N))$

The level of significance at which the null hypothesis is rejected depends on the alternative hypothesis (HA):

H_0: $d_N = d_S$

HA: (a) $d_N = d_S$ (test of neutrality).

(b) $d_N > d_S$ (positive selection).

(c) $d_N < d_S$ (purifying selection).

The probability(P) of rejecting the null hypothesis of strict-neutrality ($d_N = d_S$) in favor of the alternative hypothesis ($d_N > d_S$) is displayed in probability column of the result file generated . Values of *P* less than 0.05 are considered significant at the 5% level. substitutions per site, respectively. The variance of the difference is computed using the bootstrap method (500 replicates). [29, 30].

A similar study was conducted using 161 human house keeping genes with orthologous sequences from 3 species - Chimpanzees , Old world monkeys and Rat. There was no positive selection detected [31]. Because in this study, species closely related to humans are taken, we try to extend it by including more divergent species belonging to mammals. Species in our analysis include – *Homo sapiens*, *Pan troglodytes*, *Bos Taurus*, *Canis lupus*, *Macaca mulatta Mus musculus* and *Rattus norvegius*. In total 420 orthologous sequence pairs were tested for positive selection.

The homologous sequences were obtained from “Homologene” database at NCBI. As the files downloaded were in .txt format, these were converted to fasta format. Now these sequences were aligned using MEGA-CC. A perl program was then written for positive selection test. The results files contained a matrix with the last column of probability. A significant value of <0.05 indicated positive selection.

```
#!/usr/bin/perl -w
@files = ();
$folder = 'final_results' ;
opendir(FOLDER, $folder);
@files = readdir(FOLDER) ;
closedir(FOLDER) ;
shift @files ;
shift @files ;
$count = 0 ; # number of prob val <0.05
$count1 = 0 ; #total count of prob val

for($i = 0 ; $i < scalar @files ; $i= $i+3)
{
open (fh, "$folder/$files[$i]");
@text = <fh> ;
close fh ;
$text_as_string = join('', @text) ;
while($text_as_string =~ /\[\d\]\s*(\d.*?)\n/g)
{
$string = $1 ;
$string =~ s/\[.*\]//g ;
@array = split (' ', $string) ;
foreach $arr (@array)
{
if ($arr =~ /0\.04|0\.03|0\.02|0\.01|0\.00/g)
{
++$count;
}
++$count1 ;
}
print $string ;
@array = ();

}
}
print "\n", "number of prob val <0.05 = ", $count, "\n"  ;
print "total count of prob val = ", $count1 ;
$per = $count/$count1 *100 ;
```

```
print "\n", "percentage of positive selection of genes= ", $per ;
exit ;
```

3. RESULT AND DISCUSSIONS

If the probability was < 0.05 then alternative hypothesis was tested; HA: (Positive selection dN > dS) was selected.

Total number of sequence pair with probability value

< 0.05 = 72

Total sequence pairs = 420

Percentage of sequence pairs showing positive selection = 17.14

Here we found that only 17.12 percent of genes in mammals have evidence of positive selection. This percentage could even be very lesser. Since Homologene database contains mRNA sequences that are converted back from protein sequences , therefore the distinction between the different codons that codes for same amino acids could not be made. Hence some account of synonymous mutations may have been missed due to which probability of positive selection may have unfairly increased.

4. CONCLUSION:

As most of the results (83%) did not reject the hypothesis to favor the alternative hypothesis – HA : dN > dS , it may be safely inferred that human genes may not be under positive selection. By this it means that non synonymous mutations that bring change in gene function in order to adapt are not under any selection force of nature as proposed by Darwin. It essentially means that formation of new amino acid is either disfavored or neutrally favored. This is not in line with Darwin theory of evolution according to which both of the species must be adapting according to their environment and in order to accomplish this their genes must be under force of gradual natural selection to positively select for variations that favor in adaptation.

Here we cannot possibly give the specific nature or mechanism of path and manner in which all species were formed from complex structure organism to simple structure organism. But all the evidences of present day indicate that the path of devolution is quite possible. From drug resistance in bacteria , adaption in laboratory mice for much desired characteristic of unregulated reproduction, several disease resistance in humans, development of drug that would cripple a particular gene, to phenomenon of adaptive pseudogenization, all show the motif of loss of genes. On the other hand, there is huge improbability calculated in formation of new stable and functional protein as well as there is not a single gene or protein that is newly formed as a part of adaption by an organism and can be given as a supportive example for Darwin theory of evolution.

Darwin easily explained the path of evolution by citing morphological evidence. But with knowledge of sequencing at genome and proteome level, several scientists have started to doubt the formation of different life forms in a manner according to theory of natural selection. Hence inquiries in origin of life and tree of life hold lots of promise. Role of Bioinformatics can be crucial in the development of the subject of evolution. There is a scope in simulation of process of randomization in mutations that happen in genome. The phenomenon of chance associated with evolution can be programmed and tested over the virtual biological data available. And with completion of GWAS (Genome Wide Association Studies) project, the SNP variation association data with disease is available. To associate SNP variation with evolution is a potential possibility. Two efforts are already made in this direction [32, 33].

Overall computational statistics and programming can play a crucial information in mining patterns and knowledge from large amount of sequence data available.

Apart from the significance of larger questions like "from where we have come?" or "who are we?" and "where we are heading to?", the studies on evolution could be hugely beneficial in the field of medicine. There can be study to find a pattern or parameters involved in loss of genes and complexity that have formed different species in order to survive. Only which kind of genes can be lost or if two particular genes have to be lost together or if there is a sequence to loss of gene?. Or if we can exploit the intra human species variation data and then point out the adaption with geography. Answer to all these questions can be useful for the field of medicine.

REFERENCES

1. Olson, M. V. (1999). When less is more: gene loss as an engine of evolutionary change. The American Journal of Human Genetics, 64(1), 18-23.

2. Tamarkin, L., Baird, C. J., & Almeida, O. F. (1985). Melatonin: a coordinating signal for mammalian reproduction?. Science, 227(4688), 714-720.

3. Kleinman, H. K., Ebihara, I., Killen, P. D., Sasaki, M., Cannon, F. B., Yamada, Y., & Martin, G. R. (1987). Genes for basement membrane proteins are coordinately expressed in differentiating F9 cells but not in normal adult murine tissues. Developmental biology, 122(2), 373-378

4. Fink, A. L., Calciano, L. J., Goto, Y., Kurotsu, T., & Palleros, D. R. (1994). Classification of acid denaturation of proteins: intermediates and unfolded states. Biochemistry, 33(41), 12504-12511.

5. Tournamille, C., Colin, Y., Cartron, J. P., & Le Van Kim, C. (1995). Disruption of a GATA motif in the Duffy gene promoter

abolishes erythroid gene expression in Duffyâ negative individuals. Nature genetics, 10(2), 224-228.

6. Olson, M. V., & Varki, A. (2003). Sequencing the chimpanzee genome: insights into human evolution and disease. Nature Reviews Genetics, 4(1), 20-28.

7. Muchmore, E. A., Diaz, S., & Varki, A. (1998). A structural difference between the cell surfaces of humans and the great apes. American journal of physical anthropology, 107(2), 187-198.

8. Wang, X., Grus, W. E., & Zhang, J. (2006). Gene losses during human origins. PLoS biology, 4(3), e52.

9. Brunet, M., Guy, F., Pilbeam, D., Mackaye, H. T., Likius, A., Ahounta, D., ... & Zollikofer, C. (2002). A new hominid from the Upper Miocene of Chad, Central Africa. Nature, 418(6894), 145-151.

10. Zhang, Z., Harrison, P. M., Liu, Y., & Gerstein, M. (2003). Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome research, 13(12), 2541-2558.

11. Torrents, D., Suyama, M., Zdobnov, E., & Bork, P. (2003). A genome-wide survey of human pseudogenes. Genome research, 13(12), 2559-2567.

12. Zhang, J. (2003). Evolution by gene duplication: an update. Trends in ecology & evolution, 18(6), 292-298.

13. Durston, K. K., Chiu, D. K., Abel, D. L., & Trevors, J. T. (2007). Theoretical Biology and Medical Modelling. Theoretical Biology and Medical Modelling, 4, 47.

14. Reidhaar-Olson, J. F., & Sauer, R. T. (1990). Functionally Acceptable Substitutions in Two -Helical Regions of Repressor, Proteins: Structure, Function, and Genetics 7, 306-316.

15. Bowie, J. U., & Sauer, R. T. (1989). Identifying determinants of folding and activity for a protein of unknown structure. Proceedings of the National Academy of Sciences, 86(7), 2152-2156.

16. Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A., & Sauer, R. T. (1990) Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitution, Science 247, 1306-1310.

17. Thirumalai, D., & Klimov, D. K. (1999). Emergence of stable and fast folding protein structures. arXiv preprint cond-mat/9910248.

18. Meyer, S. C. (2014). Darwin's doubt: The explosive origin of animal life and the case for intelligent design. HarperOne.

19. Behe, M. J. (2009). Irreducible complexity: Obstacle to Darwinian evolution. Philosophy of Biology: An Anthology, 427.

20. Macnab, R. M. (1999). The bacterial flagellum: reversible rotary propellor and type III export apparatus. Journal of bacteriology, 181(23), 7149-7153.

21. Lievre, A., Bachet, J. B., Le Corre, D., Boige, V., Landi, B., Emile, J. F., ... & Laurent-Puig, P. (2006). KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. Cancer research, 66(8), 3992-3995.

22. Blanpain, C., Libert, F., Vassart, G., & Parmentier, M. (2002). CCR5 and HIV infection. Receptors and Channels, 8(1), 19-31.

23. Stephens, J. C., Reich, D. E., Goldstein, D. B., Shin, H. D., Smith, M. W., Carrington, M., ... & Dean, M. (1998). Dating the Origin of the CCR5</i>-Î < i> 32</i> AIDS-Resistance Allele by the Coalescence of Haplotypes. The American Journal of Human Genetics, 62(6), 1507-1515.

24. Kim, H. J., Kwak, H. K., Lee, J., Yun, Y. J., Lee, J. S., Lee, M. S., ... & Lee, K. H. (2012). Patterns of pncA mutations in drug-resistant Mycobacterium tuberculosis isolated from patients in South Korea. The International Journal of Tuberculosis and Lung Disease, 16(1), 98-103.

25. Geach, T. (2014). Genetics: APOC3 mutations lower CVD risk. Nature Reviews Cardiology.

26. Jorgensen, A. B., Frikke-Schmidt, R., Nordestgaard, B. G., & TybjÃ¦rg-Hansen, A. (2014). Loss-of-Function Mutations in APOC3 and Risk of Ischemic Vascular Disease. New England Journal of Medicine.

27. Kumar, S., Stecher, G., Peterson, D., & Tamura, K. (2012). MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. Bioinformatics, 28(20), 2685-2686.

28. Tamura, K. Tamarkin, Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular biology and evolution, 28(10), 2731-2739.

29. Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0.Â Molecular Biology and Evolution30: 2725-2729

30. Nei, M., & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular biology and evolution, 3(5), 418-426.

31. Zhang, L., & Li, W. H. (2005). Human SNPs reveal no evidence of frequent positive selection. Molecular biology and evolution, 22(12), 2504-2507.

32. Cheng, F., Chen, W., Richards, E., Deng, L., & Zeng, C. (2009). SNP@ Evolution: a hierarchical database of positive selection on the human genome. BMC evolutionary biology, 9(1), 221.

33. Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A map of recent positive selection in the human genome. PLoS biology, 4(3), e72.